APPLIED MULTIVARIATE ANALYSIS

B. A. CHANSARKAR

Himalaya Publishing House

ABOUT THE BOOK

This book deals with applications of multivariate techniques, hetherto neglected due to limitations of data collection and non-availability of computing facilities. Seven important multivariate techniques have been elucidated with applications in various fields both educational and social/industrial. The applications of multivariate techniques has now become a common feature of national planning, business modelling psychometric analysis and all aspects of research.

The book will be of immense value to all students of Statistics, Economics, Business Administration and all those engaged in research in higher education and in Industry.

PRICE Rs. 70-00

Applied Multivariate Analysis

B.A. CHANSARKAR

FOREWORD BY

M.A. Chansarkar
Vice-Chancellor
Nagpur University

HPH
Himalaya Publishing House
BOMBAY ✶ DELHI ✶ NAGPUR

FIRST EDITION: OCTOBER, 1987

Published by: Mrs. Meena Pandey for
HIMALAYA PUBLISHING HOUSE

Head Office:

"Ramdoot", Dr. Bhalerao Marg,
Girgaon (Kelewadi),
Bombay-400004

Branch Office:

"Pooja Apartment", 4-B, Murarilal Street,
Ansari Road, Darya Gunj,
New Delhi-110002

Printed by:

D.B. Sharma & Sons,
129/4, Sane Guruji Marg,
Bombay-400011

NAGPUR UNIVERSITY

DR. M. A. CHANSARKAR
VICE CHANCELLOR

FOREWORD

My brother Dr. B.A. Chansarkar from Middlesex Business School, U.K. who is an eminent scholar, was a Visiting Professor in Nagpur University in the Post-Graduate Teaching Department of Statistics during 1980-81. During this period, on the request from the University he conducted five enlightening staff seminars on ''Applications of Multivariate Techniques in various Situations'' which were highly appreciated by one and all. This book is an enlarged and improved version of the same.

Indian Statisticians have been playing a dominant role in the development of statistical theory, especially in the field of Multivariate Analysis. Over the last thirty years, Multivariate Analysis has been playing an important role in all aspects of statistical interpretation. This has now been further strengthened by the development and usage of computors as a part of educational technology. With the developing computer technology, extensive use is made which was hitherto difficult due to time consuming manual calculations. With easy accessibility to advance computer packages, multivariate techniques are being commonly used in all disciplines.

It has become an essential research tool to tackle problems of complicated nature-both scientific and social. Modern facilities for extensive data collection and storing are easily available and this has further encouraged the use of multivariate techniques. The applications of multivariate techniques has now become a common feature of national planning, business modelling, psychometric analysis and all aspects of research, both in educational institutions and industry.

The present book by Dr. B.A. Chansarkar is one of the first attempts of its kind in this highly specialised field. I am sure this publication will go a long way in helping students, teachers as well as businessmen and industrialists, in developing an indepth understanding of use of multivariate techniques, as they have immensely helped the Faculty Members of Nagpur University.

Nagpur
September 1987

M.A. CHANSARKAR

PREFACE

This book on Applied Multivariate Analysis is based on a series of staff seminars conducted by me at the Post-Graduate Department of Statistics at Nagpur University, Nagpur (India) while I went there as a Visiting Professor. During this period, I having experienced the method of teaching and realising the theoretical approach to the subject in India, felt the need for this kind of a book elucidating the applications of these techniques.

This objective is achieved through this book to some extent by concentrating on only those techniques which are commonly used. I hope, the book will be of immense value to the students of Statistics, Economics and all those engaged in research either for higher degree or in Industry.

My sincere thanks are due to Nagpur University and its Faculty members because of whose generous help and assistance the Visiting Professorship & seminars were made possible.

Ordinarily, my brother, Dr. M.A. Chansarkar, Vice-Chancellor, Nagpur University would neither expect nor would like to be thanked for writing the Foreword to this publication. However, I will be failing in my duty, if I would not express my gratitude for the same.

B.A. CHANSARKAR

Enfield, UK
August 1987

CONTENTS

1. Introduction

1.1 Multivariate statistical methods are widely used in the last decade or two because they make it possible to encompass all the data from an investigation in one analysis. This approach results in a clearer, better organised account of the investigation than the piecemeal analysis of the parts of the data which are often observed in behavioural studies. It also permits more realistic probabilistic statements in hypothesis testing and interval estimation than do separate analysis. Such a comprehensive approach necessitates a foresight and organising ability on the part of the researcher and requires of successful investigation a proper use of statistical models to focus thinking and express hypotheses succinctly. Such models, in behavioural studies, tend to be multivariate since many response variables are observed simultaneously.

The main contributors for the development of multivariate methods have been H. Hotellings, R.A. Fisher, S.S. Wilks, S.N. Roy, M.S. Bartlett, C.R. Rao, and T.W. Anderson. The availability of computers to perform the laborious computations required in multivariate analysis when the number of variables exceeds 2 or 3, has helped in development of statistical theory and its wide applications in various fields, such as marketing, industrial and business analysis, attitude measurements, agriculture and behavioural sciences.

The type of analysis (see Appendix A) one can perform depends heavily on the available or collected information. It is in this context the type of measurement and its realiability is quite important. There are basically two main types of (measurements) scales-a) metric b) non-metric. These types are determined, both by the empirical operations involved in the process of measuring and the mathematical properties of the scale.

1.2 The four classes of measurements (1) are nominal, ordinal, interval and ratio. The first two are non-metric and other two are metric.

Nominal (Classificatory):
Crudest form of measurement.
Measurement at weakest level.
Statistic relevant: Mode

Ordinal (Ranking):
Rank order.
Statistic relevant: Median
Percentile
Spearman's r

Interval (Cardinal):
Distances between any two numbers on the scale are of known size.
Unit of measurement arbitrary.
Quantitative scale.
Statistic relevant: Mean
Standard deviation
Pearsons r
Multiple R

Ratio:
Zero as origin.
Rarely used in marketing conditions.
Statistic relevant: Geometric Mean
Coefficient of variation

1.3 Sample selection and the material (universe) from which the sample is drawn is quite important. The sample should be drawn randomly, independently from a large population. The reliability of the sample estimate is based on the standard error of the corresponding estimate, e.g., standard error of sample mean $S_{\bar{x}} = \frac{S}{\sqrt{n}}$.

The most commonly used measure of location is the arithmetic mean $\bar{x} = \frac{\Sigma x}{n}$, though certain situations may demand use of other measures (median, mode). The squareroot of variance (s.d. $s_x = \sqrt{\frac{1}{n} \sum (x-\bar{x})^2}$ and covariance $s_{xy} = \sqrt{\frac{1}{n}\left[\sum (x-\bar{x})(y-\bar{y})\right]}$ are the most commonly used measures of dispersion, though the mean deviation $\frac{\sum |x|}{n}$ or absolut mean deviation may be suitable in some situations. Wit several measurements being recorded on an individual different scales (units) are used and the variance — covarianc matrix is dependent on the units of measurement. Therefore in many situations Pearson's product correlation coefficient

$$r_{xy} = \frac{\sum (x-\bar{x})\ (y-\bar{y})}{\sqrt{\sum (x-\bar{x})^2 \sum (y-\bar{y})^2}}$$

are used. r can vary onl between ± 1. When $r = \pm 1$, all points lie exactly o a straight line and there is perfect relationship and when r = c there is no *linear* relationship. This property is very useful in ir terpreting r. Further r^2, the coefficient of determination, i useful for interpretation purposes as it tells us the percentage c the variation in y is explained by the variation in x (and vice versa). External validation is necessary to know whether th relationship is casual one or not.

REFERENCES

(1) Siegel S 'Non-parametric statistics for Behavioural Sciences' McGraw Hill London 1956

(2) Moser C A & Kalton G 'Survey Method in Social Investigations' Heinemann Educational London 1976

(3) Taylor, MB 'Ordinal and Interval Scaling' Journal of the Market Research Society Volume 25, No 4, October 1983.

2. Regression Analysis

2.1 Dependence analysis with one variable being dependent and the rest independent.

Assumption (1): One of the standard assumptions of regression analysis is that the model which describes the data is linear. The relationship (as seen from the scattergram) may be non-linear. There are several non-linear models which by appropriate transformations[1] can be made linear (see Appendix B).

Assumption (2): The other standard assumption is constancy of error variance. When the error variance is not constant over all the observations, the error is said to be heteroscedastic. There are transformations[2] which stabilise the variance and also have the effect of making the distribution of the transformed variable closer to the normal distribution (see Appendix C).

The data consist of n observations on a dependent variable y and p independent (explanatory) variables x_1, x_2, ...x_p

Let

$$x = \begin{bmatrix} x_{01} & x_{11} \ldots\ldots x_{p1} \\ x_{02} & x_{12} \ldots\ldots x_{p2} \\ \cdot & \cdot \qquad\quad \cdot \\ \cdot & \cdot \qquad\quad \cdot \\ \cdot & \cdot \qquad\quad \cdot \\ x_{0n} & x_{12} \ldots\ldots x_{pn} \end{bmatrix} \quad y = \begin{bmatrix} y_1 \\ y_2 \\ \cdot \\ \cdot \\ \cdot \\ y_n \end{bmatrix}, u = \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ \cdot \\ u_n \end{bmatrix} \text{ and } \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{bmatrix}$$

The x's are non-stochastic, i.e. values of x's are fixed and x's are measured without error.

The linear model representing the data is

$y = x\,\beta + u$ where $x_{0i} = 1$ for all i.

The assumptions made about u for the least squares estimator are

$$E(u) = 0, \; Var(u) = E(uu') = \sigma^2 In$$

i.e., u's are independent, have zero mean and constant variance.

This implies $E(y) = x\beta$.

The least square estimate of ß is obtained by minimising the sums of squares of deviations from their expected values.

$S(\beta) = u'u = (y-x\beta)' \; (y-x\beta)$.

Minimum of S(ß) leads to the system of equations

$(x'x)b = x'y$ (called Normal equations)

$\therefore b = (x'x)^{-1}x'y \quad$ iff $|x'x| \neq 0$.

The vector of predicted values of $\hat{y}$ corresponding to y is $\hat{y} = xb$.

The vector of residuals is $e = y-\hat{y} = y-xb$.

2.2 The properties of the least squares estimators are

(1) b is an unbiased estimator of ß, with variance-covariance matrix V(b) which is

$V(b) = E(b-\beta)(b-\beta)' = \sigma^2(x'x)^{-1} = \sigma^2{}_c$

where $c = (x'x)^{-1}$ and $E(b) = \beta$

(2) The unbiased estimator of σ^2 is S2 where

$$S^2 = \frac{e'e}{n-p-1} = \frac{(y-\hat{y})'(y-\hat{y})}{n-p-1} = \frac{y'y-b'x'y}{n-p-1}$$

with added assumption ui's are normally distributed, we have

(3) Vector b has p variate normal distribution with mean ß and variance $\sigma^2{}_c$. The marginal distribution of bi is normal with mean β_i and variance $\sigma^2{}_{cii}$ where c_{ii} is the ith diagonal element of c.

(4) The quantity $\quad W = \dfrac{e'e}{\sigma^2} \quad$ is $\quad \chi^2{}_{n-p-1}$

(5) b and S^2 are distributed independently of each other.

2.3 The adequacy of the linear model is made by the multiple correlation coefficient R^2.

$$R^2 = 1 - \frac{\Sigma(y_i - \hat{y})^2}{\Sigma(y_i - \bar{y})^2} = 1 - \frac{e'e\ /n}{\Sigma(y_i - \bar{y})^2/n} = 1 - \frac{\hat{\sigma}^2}{S_y^2}$$

The estimates will be unbiased estimators if corrected for the degrees of freedom.

$$\text{So} \quad \bar{R}^2 = 1 - \frac{\hat{\sigma}^2/n-p}{S^2_{y/p-1}} = 1 - (1-R^2)\frac{n-1}{n-p}$$

$\bar{R}^2$ enables one to compare separate regression equations with different number of pre-determined variables X in the equations. It is this aspect of total variation explained enables one to choose the variables to be included in the model.

2.4 Forward Stepwise regression analysis or Backward Elimination procedure:

In forward stepwise regression analysis, the first variable included in the equation is the one with the highest correlation coefficient. If b is significantly different from zero, the first variable is retained and search is made for the second variable. Variable with the highest simple correlation coefficient with the residuals from step one is then included. If b for the second variable is significantly different from zero, it is retained and search is made for the third one in the same manner. The cut off comes when the last variable entering has non-significant regression coefficient or all the variables are included in the equation. The backward elimination procedure is exactly the reverse. You start with full equation with all the variables included and successively drop one variable at a time. The variables are eliminated on the basis of the contribution to the reduction of error sums of squares.

Dummy variables are used in a variety of ways and may be considered whenever there are qualitative factors affecting a relationship, e.g., sex, marital status, political affiliation etc. One assigns a value zero or one to the variable according to whether the respondent has no characteristic or possesses the characteristic under investigation.

2.5 Application of Regression Analysis.

In order to assess the contribution of different factors affecting the foodgrains production in India[3], linear regression analysis was performed over a period of 26 years 1951-1977, using the following set of variables.

y = Total production of foodgrains (Million Tonnes)

x_1 = Public sector outlay at constant 1961-62 prices on Agriculture and related sectors (Rs. Crores)

x_2 = Weather conditions 0 when normal
1 otherwise (draughts/
and or floods)

x_3 = Availability of fertilisers (000 Tonnes)

x_4 = Gross area irrigated under foodgrains (M. Hectares)

x_5 = Gross area Unirrigated under foodgrains (M. Hectares)

x_6 = Net Imports of foodgrains (M. Tonnes)
Proxy for 'gap between requirement and availability'.

x_7 = Wholesale price index of foodgrains (1961-62 = 100) Derived Series.

The equation with all variables was

$$y = -105.28 + \underset{(0.01)}{0.005\,x_1} - \underset{(1.53)}{2.77\,x_2} + \underset{(0.01)}{0.004\,x_3} + \underset{(0.95)}{2.23\,x_4}$$

$$+ \underset{(0.23)}{1.45\,x_5} - \underset{(0.47)}{0.85\,x_6} + \underset{(0.06)}{0.01\,x_7}$$

Figures in brackets are standard errors of coefficients.

$R^2 = 0.97$ $\bar{R}^2 = 0.96$

Forward Stepwise regression analysis was performed to judge the importance of variables to include in the model. The final equation being

$$y = -115.19 - \underset{\substack{(1.51)\\ \text{weather}}}{3.03\,x_2} + \underset{\substack{(0.20)\\ \text{area}\\ \text{irrigated}}}{3.33\,x_4} + \underset{\substack{(0.21)\\ \text{area}\\ \text{unirrigated}}}{1.29\,x_5} - \underset{\substack{(0.29)\\ \text{imports}}}{0.83\,x_6}$$

Figures in brackets are standard errors of coefficients.

$R^2 = 0.96$ $\bar{R}^2 = 0.96$

It is interesting to note that the two variables x_4 and x_5—area—contribute 90% of the total variation.

2.6 The problem of correlated errors *(Autocorrelation)*

When the observations have a natural sequential order, the correlation is referred to as autocorrelation. The presence of autocorrelation has the following effects:

(1) Least squares estimates are unbiased but not efficient in the sense they no longer have minimum variance.

(2) The estimates of σ^2 and the standard errors of the regression coefficients may be seriously understated, giving spurious impression of accuracy.

(3) The confidence intervals and various tests of significance commonly used would be no longer valid.

Two types of autocorrelation can occur in practice. Firstly, autocorrelation in appearance due to omission of a variable that should be in. Once the variable is uncovered the problem is resolved. Secondly, it could be pure autocorrelation. Correction involves transformation of data.(4)

Detection of Autocorrelation by Durbin-Watson statistic. The amount of autocorrelation that exists in the residuals is measured by the Durbin-Watson statistic.

Errors constitute first order autoregression series.

$$U_t = \rho U_{t-1} + \epsilon_t \qquad |\rho|<1$$

ϵ_t — Independent and Normally distributed with zero mean and constant variance.

$$\text{Let } d = \frac{\sum_{2}^{n} (\rho_t - \rho_{t-1})^2}{\sum_{1}^{n} \rho_t^2}$$ be the test statistic.

for Ho: $\rho = 0$

H_1: $\rho > 0$

We estimate parameter ρ by r where

$$r = \frac{\sum_{2}^{n} e_t \, e_t - 1}{\sum_{1}^{n} e_t^2}$$ approximately.

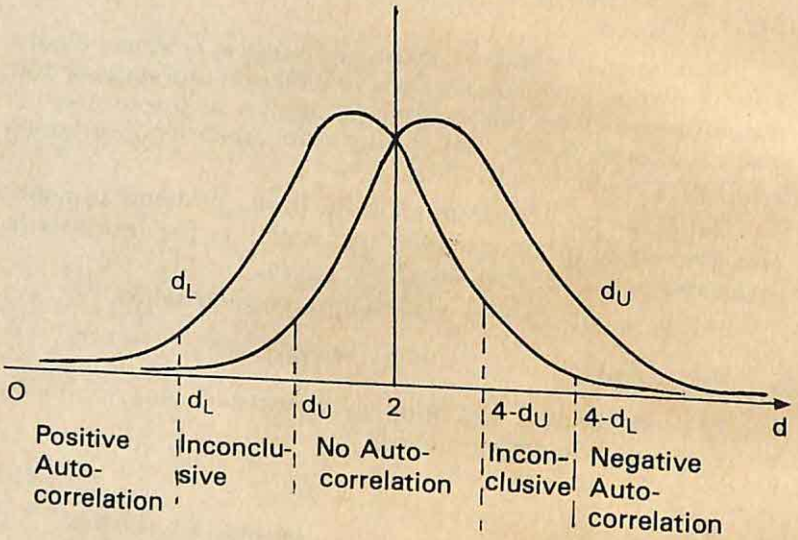
$d \simeq 2\,(1 - r)$.

d changes between 0 and 4.

$d \simeq 0$ if $r = 1$
$d = 2$ if $r = 0$
$d = 4$ if $r = -1$

deviation of d from numerical value 2 indicates autocorrelation.

Rule: if $d < d_L$ Reject H_O
$d > d_U$ do not reject H_O
$d_U < d < d_L$ the test is inconclusive.



Example

Consumer expenditure (y_t) and money stock (x_t) in U.S.A., 1952 to 1956, quarterly data, Units of measurement billion current dollars.

$$y = -154.700 + 2.300\, x_t$$
$$(19.85) \qquad (0.115)$$

(Figures in brackets are standard errors of the estimates)

$R^2 = 0.955$

$d = 0.328$ $\qquad d_L = 1.28$ for $n = 20$

d significant at 1% level.

Cochrane-Orcutt[5] have suggested the following method using transformation to remove autocorrelation:

Procedure is: use $y_t - \rho y_{t-1}$ and $x_t - \rho x_{t-1}$ instead of y_t and x_t.

Estimate ρ by first ordinary least squares and use the errors to get $\hat{\rho}$.

For the USA data above, $\hat{\rho} = 0.874$

using this $y_t = -324.44 + 2.758\, x_t$

(0.44)

(Figures in brackets is standard errors of the estimate)

$d = 1.607$ accepted at 5% level of significance since $d_U = 1.49$.

2.7 Multicollinearity: The phenomenon of mutual linear dependence between the explanatory variable x_i is called multicollinearity[6]. When there is complete absence of linear relationship among the explanatory variables, they are said to be orthogonal. It affects statistical inference[7] and forecasting[8].

Indication of multicollinearity that appear as instability in the estimated coefficients are as follows:

(1) Large changes in the estimated coefficients when a variable is added or deleted.

(2) Large changes in the estimated coefficients when a data point is altered or dropped.

Once the residual plots indicate that the model has been satisfactorily specified, multicollinearity may be present if

(3) The algebraic signs of estimated coefficients do not conform to prior expectations or

(4) Coefficients of variables that are expected to be important have large standard errors.

To overcome multicollinearity use Principal Components (see Chapter 4 for further details).

REFERENCES

(1) Daniel C & Wood F S 'Fitting Equations to Data' Wiley New York 1971

(2) Kendall M G & Stuart A 'The Advanced Theory of Statistics' Vol. 3 Charles Griffin London 1968

(3) Chansarkar B A 'Factors Affecting Foodgrains Production in India' Indian Society for Agricultural Statistics, 34th Annual Conference Lucknow 1980

(4) Johnston J 'Econometric Methods' McGraw Hill New York 1972
Kmenta J 'Elements of Econometrics' MacMillan New York 1971

(5) Cochrane D & Orcutt G H 'Application of least squares regression to relationships containing autocorrelated error terms' Journal of American Statistical Association 44 32-61 1949

(6) Ferror D E & Glauber R 'Multicollinearity in regression analysis, the problem of revisited' Review of Economic & Statistics 92-107 February 1967

(7) Coleman Mosteller et al 'Research on equal opportunities in public education in USA'

(8) Mulinvaud E 'Statistical Methods of Econometrics' Rand McNally Chicago 1968

Other useful references for regression analysis:

(a) Frank R E 'Use of Transformations' Journal of Marketing Research August 1966

(b) Schildernick J H F 'Regression and Factor Analysis applied in Econometrics' Martinus Nijhoft Social Sciences Division Leiden 1977

(c) Chatterjee S & Price B 'Rregression analysis by example' Wiley New York 1977

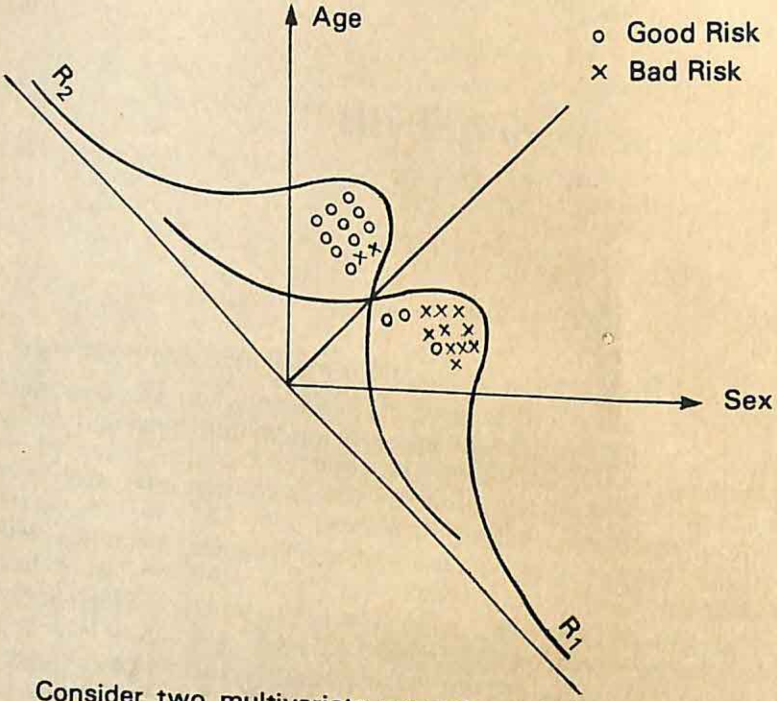
3. Discriminant Analysis

3.1 Dependence analysis but more than one dependent.

The problem of classification arises when an investigator makes a number of measurements on an individual and wishes to classify the individual to either one or the other of the groups. Consider the situation where you are interested in attempting to discriminate new Ford and Chevrolet buyers[1]. The characteristics used are personality needs, socio-economic variables and a combination of both. Similarly as a Bank Manager you want to decide whether to advance a loan or not? Discriminant analysis is useful in situations where a total sample is divided into known groups based on some classificatory variable and the researcher is interested in understanding the group differences or in predicting correct belonging to a group of a new sample based on the information on a set of predictor variables. Other examples of discriminatory analysis being among listeners who sent for a programme guide from those who did not[2], and effective new product decisions for supermarkets[3] and types of holders of savings accounts[4].

3.2 In constructing the procedure of classification, it is desired to minimise the probability of misclassification. The space is to be divided into regions R_1 and R_2 such that expected loss is as small as possible. The Bayes procedure is usually used — it is a minimax procedure if the maximum expected loss is a minimum[5].

A discriminant function is a linear function of the set of observations weighted by the inverse of the variance-covariance matrix. The linear function has the greatest variance between samples relative to the variance within samples. It is thus analogous to one way classification of analysis of variance[6].



Consider two multivariate normal populations with equal covariance matrices, namely, $N(\mu^{(1)}, \Sigma)$ and $N(\mu^{(2)}, \Sigma)$, where $\mu^{(i)} = (\mu_1^{(i)}, \dots\dots \mu_p^{(i)})$, is the vector of means of the ith population, i= 1,2. and Σ is the variance-covariance matrix of each population. Further, let x be an observation. We wish to classify x to either of the population.

The discriminant function is $x\ \Sigma^{-1}\ (\mu^{(1)} - \mu^{(2)})$

and if the two populations are equally likely and costs of misclassification are equal, then if the discriminant function has value greater than a scalar $(= \frac{1}{2}\ (\mu(1) + \mu(2))'\Sigma{-1}\ (\mu(1) - \mu(2))$ then the observation x belongs to the first population, otherwise to the second population.

$$x' \Sigma^{-1} (\bar{x}^{(1)} - \bar{x}^{(2)}) - 1/2\ (\bar{x}^{(1)} + \bar{x}^{(2)})\ S^{-1}\ (\bar{x}^{(1)} - \bar{x}^{(2)}) \geq \log K$$

where $(n_1 + n_2 - 2)\ S = \sum_{1}^{n_1} (x_a^{(1)} - \bar{x}^{(1)})\ (x_a^{(1)} - \bar{x}^{(1)})' + (x_a^{(2)} - \bar{x}^{(2)})\ (x_a^{(2)} - \bar{x}^{(2)})'$

3.3 Once the discriminant function is obtained, the function could be tested for significance. Then, the exact classification of all the individuals in the sample is worked out. The proportion of correct classification is then compared against what could have been predicted by chance without any knowledge of the scores on the predictor variables[7]. However, to avoid biases, it is more appropriate to validate the analysis by using the discriminant weights on another sample of individuals, though sometimes it is preferable to split the sample into half, using one half for analysis and the other half for validation.

In a certain bank, in order to advance loans to prospective applicants the characteristics of previous 52 clients chosen randomly have been studied over a period of two years and in the given localities. It is known that of these 25 were good risk (returned the loan as per the terms) and 27 were bad risk. Information on the following characteristics was studied:

x_1 = Sex O = Male, 1 = Female
x_2 = Number of years of service
x_3 = Number of children
x_4 = Net weekly income adjusted for taxes
x_5 = Average weekly rent/mortgage.

Preliminary analysis indicates variables x_1 and x_3 are not significantly different and overall discrimination is not satisfactory. Further run with only 3 remaining variables x_2, x_4 and x_5 gave the following results:

$$\begin{array}{ccc} \multicolumn{2}{c}{\text{Means}} & \\ \text{Good Risk} & \text{Bad Risk} & \text{Matrix of corrected sums of squares and products} \\ \begin{bmatrix} 7.40 \\ 87.84 \\ 36.44 \end{bmatrix} & \begin{bmatrix} 5.95 \\ 35.67 \\ 21.93 \end{bmatrix} & \begin{bmatrix} 20.96 & 421.92 & 148.24 \\ & 88897.36 & 9135.09 \\ & & 6090.12 \end{bmatrix} \end{array}$$

The discriminant function is

$$3.032\, x_2 + 0.012\, x_4 + 0.032\, x_5 \geq 21.90$$

Then the Population is of good risk.

An individual with 6 years of service, weekly net income of £31.0 and mortgage of £23 week is a bad risk (as the discriminant function has the value of 19.91) and will not be advanced the loan.

The classification of all the individuals gave the following result:

		Observed		
		Good	Bad	
Expected	Good	23	6	29
(By discriminant analysis)	Bad	2	21	23
		25	27	52

The Fishers exact test[6] has probability less than 0.05 and hence the null hypothesis of independence is rejected.

REFERENCES

(1) Evans F B 'Psychological & Objective Factors in Brand Choice' Journal of Business Vol. 31 October 1959

(2) Massy W F 'Discriminant Analysis of Audience Characteristics' Journal of Advertising Research Vol. 5 No. 1 March 1965

(3) Doyle P & Weinberg C B 'Effective New Product Decisions for Supermarkets' Operations Research Quarterly June 1973

(4) Claycamp H J 'Characteristics of Owners of Thrift Deposits in Commercial Banks and Savings and Loan Associations' Journal of Marketing Research Vol. 2 May 1965

(5) Anderson T W 'Introduction to Multivariate Statistical Analysis' Wiley New York 1958

(6) Fisher R A 'Statistical Methods for Research Workers' Haftner Publishing Co New York 1958

(7) Morrison D G 'On the Interpretation of Discriminant Analysis' Journal of Marketing Research May 1969

4. Principal Component Analysis

4.1 Interdependence analysis with metric inputs.

In the analysis of interdependence, there is no one variable or variable subset that is the focus of study that differs in importance from the others. A variable or set of variables is not to be predicted from the others or explained by them. The goal, rather, is to give meaning to a set of variables or objects.

When variables are related, they can be made orthogonal (and therefore completely independent) by the method of principal component analysis.

Every linear regression model can be restated in terms of a set of orthogonal explanatory variables. They are referred to as the principal components of the explanatory set of variables.

Principal components are linear combinations of variables which have special properties in terms of variances, e.g. the first principal component is the normalised linear combination (i.e. the sums of squares of coefficients being one) with maximum variance. The principal components turn out to be the characteristic vectors of the covariance matrix(1).

Suppose X has p variables (measurements) with covariance matrix Σ. The actual distribution of X is irrelevant except for the covariance matrix, however, if X is normally distributed more meaning can be given to the principal components.

Let C be a p component column vector such that $C'C = I$
The variance of $C'X$ is

$$E(C'X)^2 = E(C'XX'C) = C'\Sigma C \qquad (1)$$

To determine the normalised linear combination $C'X$ with maximum variance, we must find a vector C satisfying $C'C = I$ which maximises (1).

Let $\Phi = C'\Sigma C - \lambda\,(C'C - I)$ where λ is the lagrange multiplier. The vector of partial derivatives $\frac{\delta\Phi}{\delta C_i}$ is

$$\frac{\delta\Phi}{\delta_c} = 2\Sigma C - 2\,\lambda C \tag{2}$$

$$\text{Then } (\Sigma - \lambda\, I)\, C = 0 \tag{3}$$

$$\text{and since } C'C = I, \qquad |\Sigma - \lambda I| = 0 \tag{4}$$

(4) has p roots, $\lambda_1 \geq \lambda_2 \geq \ldots\ldots \geq \lambda_p \geq 0$

Consider, the orthogonal transformation $U = C'X$

The $U_1, U_2, \ldots\ldots U_p$, the elements of U are the principal components. U_1 corresponds to the maximum eigen value λ_1 and is called the first component, U_2 is the second and so on. The total variation remains the same, even after transformation from X to U.

4.2 Properties of the principal components

(1) Number of components same as number of variables, p in our case.

(2) The new variables (i.e. components) are uncorrelated.

(3) Components come out in order of importance in a descending order. (Importance measured by variation explained).

The principal components are based on covariance matrix Σ. Often, the sample has mixed variables which are measured in different units. For example, the psychologist standardises the various tests first. In such situations there is justification for standardising throughout and converting the covariance matrix Σ to a correlation matrix R. The disadvantage is that if the original variables are p - variate normal, the now standardised observations are not, or rather they are approximately normal as n becomes large. Another difficulty is that, this device distorts the original measurements which, for example, took due account of the fact the head is a portion of a much larger body. In the standardised scale the head and the body length are on equal footing. The principal components are obtained from the eigenvalues of R. Some information contained in Σ is lost but the lost information is kept to a minimum(2).

4.3 The two main uses of principal component analysis are

(1) Getting independent explanatory set of variables(3).
(2) Reduction in number of explanatory variables(4),(5).

Example

Consider the data concerning the import activity of the French economy(6) over a period of 1949 to 1966.

y = Import
x_1 = Domestic Production
x_2 = Stock formation
x_3 = Domestic Consumption.

All variables in Milliards of French Francs.

Multiple Regression analysis gives

$$\underset{(4.125)}{y = 19.730} + \underset{(0.187)}{0.032x_1} + \underset{(0.322)}{0.414x_2} + \underset{(0.285)}{0.243x_3}$$

(Figures in brackets are standard errors of coefficients)

$R^2 = 0.973 \qquad n = 18$

$F_{3,14} = 168.45$

The model is not well specified and it is known this is partially due to France's trade with EEC in 1960.

The correlation matrix R being

$$R = \begin{array}{c} \begin{array}{ccc} X_1 & X_2 & X_3 \end{array} \\ \begin{bmatrix} 1 & 0.026 & 0.997 \\ & 1 & 0.036 \\ & & 1.00 \end{bmatrix} \end{array} \qquad X_i = \frac{x_i - \bar{x}}{\sigma_x}$$

The principal components are

$$u_1 = 0.7063X_1 + 0.0435X_2 + 0.7065X_3$$
$$u_2 = -0.0357X_1 + 0.990X_2 - 0.0258X_3$$
$$u_3 = -0.7070X_1 - 0.0070X_2 + 0.7072X_3$$

$$\begin{array}{c} \begin{array}{ccc} u_1 & u_2 & u_3 \end{array} \\ \begin{bmatrix} \delta_1 & 0 & 0 \\ & \delta_2 & 0 \\ & & \delta_3 \end{bmatrix} \end{array}$$

δ's are variances of u_i's and eigenvalues of R.

If δ_i's are all equal unity, the original variables are orthogonal.

If $\delta_i = 0$ exactly, there is perfect linear relationship among original variables — an extreme case of multicollinearity.

If one of the δ_i is much smaller than the other (and near zero), multicollinearity is present.

$\delta_1 = 1.999$, $\delta_2 = 0.998$, $\delta_3 = 0.003$

Since δ's are variances of the principal components, if δ is approximately zero, the corresponding component is approximately equal to zero.

Here $\delta_3 = 0.003 \simeq 0$ u_3 is constant and mean of u_3 is zero.

i.e., $u_3 = -0.7070X_1 - 0.0070X_2 + 0.7072X_3 = 0$

$\therefore X_3 = X_1$

This is consistent with $\mathbf{r}_{x_1x_3} = 0.997$.

Estimate ß$_1$ and ß$_3$ by doing regression.

$y = ß_2X_2 + (ß_1 = ß_3)\, X_4$

where $X_4 = X_1 + X_3$

$y = 0.612X_2 + 0.086X_4 - 9.007$
(0.109) 0.003)

(Figures in brackets are standard errors of the estimates)

$R^2 = 0.987$

Both regression coefficients positive and significant.

The final equation then is

$y = 0.086X_1 + 0.612X_2 + 0.086X_3 - 9.007$

REFERENCES

(1) Lawley D N & Maxwell A E 'Factor Analysis as a Statistical Method' Butterworth London 1971

(2) Cooley W W & Lohnes R R 'Multivariate Procedures for Behavioural Sciences' John Wiley New York 1962

(3) Holmes C 'Construction and Stratification of a Sampling Frame of Primary Sampling Units' The Statistician Vol 19 London 1969

(4) Massey W F 'TV Ownership in 1950' in Quantitative Techniques in Marketing Analysis' by Frank, Kuehn & Massey 1962

(5) Prakash A & Agarwal D K 'Use of Principal Component Analysis for estimation of production of Berseem (Fodder) Crop' Indian Journal of Agricultural Statistics, Vol XXVII No. 2 December 1975

(6) Mulinvaud E 'Statistical Methods of Econometrics', Rand McNally, Chicago, 1968.

5. Factor Analysis

5.1 Factor Analysis is concerned with resolution of set of descriptive variables in terms of a small number of categories or factors. This is achieved by analysis of intercorrelations of the variables. The factors this generates convey all the essential information of the original set of variables.

Factor analysis differs from principal component analysis in two respects. First, the variables are assumed to be analysable into a small sét of factors and an error term. This error term does not appear in principal components. A position of ϵ_j associated with variable j is often identified as being a specific factor. The remaining factors are then termed common factors. The second difference involves the process of rotative factors to new orthogonal or even non-orthogonal axes, if such a rotation will improve the interpretability of the resulting factors.

Suppose $X = \Lambda f + \mu + \epsilon$

where f = m-component vector of (non-observable) factors

$\underset{\sim}{\mu}$ = fixed vector of means

$\underset{\sim}{\epsilon}$ = vector of (non-observable) errors

Λ = (pxm) matrix of factor-loadings ($m < p$)

where f is random

we assume $E(f) = O$, $E(\epsilon) = O$,

$E(ff') = M$, $E(\epsilon\epsilon') = \psi$ and diagonal and

$E(f\epsilon') = O$

$E(X) = \mu$

and $\text{Cov}(X) = E(X-\mu)(X-\mu)' = \Lambda M \Lambda' + \psi$

Thus, $\Sigma = \Gamma + \Psi$

$$\text{Correaltion Matrix} = \begin{bmatrix} h_1^2 & r_{12} & r_{13} \text{---} r_{1n} \\ r_{21} & h_2^2 & r_{23} \text{---} r_{2n} \\ \\ r_{21} & r_{n2} & r_{n3} \text{---} h_n^2 \end{bmatrix}$$

where h_j^2 are communalities (unknown), and are obtained with the help of inter-correlations. As a first estimate one can take square of multiple correlation coefficient of that variable with all the other original x variable.

Taking all variables in standard form (0 mean and unit variance)

$$Z_J = a_{jl}F_l + a_{j2}F_2 + \text{............} + a_{jm}Fm + a_j\,\epsilon_j$$

The sums of square of common factors coefficients is

$$h_j^2 = \sum_{s=1}^{m} a2_{js}$$

Placing ones in the diagonal of the matrix presumes that the variance is partitioned only among common factors obtained by doing principal component analysis. It is this concern which has led to the use of other values[1] than one in the diagonal.

5.2 The choice of m — the number of factors — is quite important. When a large number of unorganised set of variables is factored, the analysis will extract the largest and the most interesting combinations of variables first and then proceed to smaller combinations. Carrying the analysis too far (more than 7 factors) has two penalties. It is exceedingly wasteful on computer time and it obscures the meaning of the findings because it affects the rotation adversely.

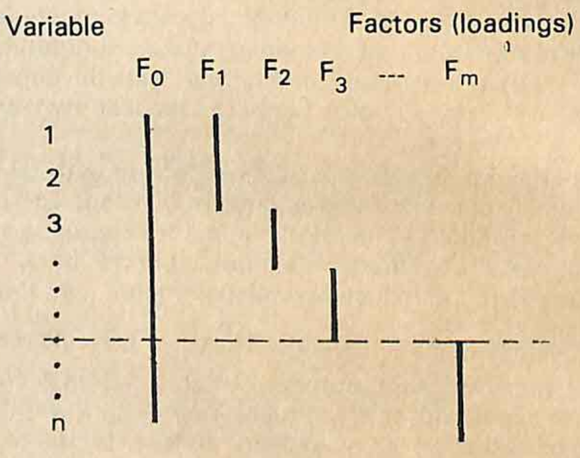
Four stopping criteria may be employed. When the analyst already knows enough about his data so that he knows how many factors are actually there, he can have the analysis stopped after that number of factors has been extracted. Secondly, if he has a clear idea in advance about the amount of variance the factors can explain, he can stop when that criterion is reached. Most commonly, however, if he does not know very much about his data to begin with, he will want to keep factoring until the factors get small and meaningless. The third one is an incremental approach. After a first set of factors has explained a large percentage of variation, say 70 per cent, if the next factor adds only a small percentage to the total variance, say less than five per-cent, it may be discarded and we could stop factoring. The final criterion is most objective. It states that all factors whose eigenvalues are greater than one when a correlation matrix is factored can be considered as significant and meaningful factors. The last two criteria are statistical in nature.

5.3 The two commonly used methods of extracting factors are Principal factor solution (Centroid Method) and Maximum likelihood solution.(2) The principal factor solution is simply the application of the mathematics of principal component analysis (Chapter 4.1) to the reduced correlation matrix, i.e., the one with communalities (h_j^2) in place of one's. The principal factor solution has the important property of the principal component analysis inasmuch that it produces factors in the order of the amount of variation they explain. It also tends to produce bipolar factors with some high negative loadings as well as some high positive loadings. Nowdays, the method of maximum likelihood solution is becoming increasingly available. This is an efficient method of extracting factors which does not require advance knowledge of the communalities (h_j^2) but does need to be told the number of factors to be produced by the analysis.

To get over the problem that analysis gives more than one set of factor loadings, methods have been developed to produce solutions which are unique in a certain sense. These methods are known as rotations. It is important to note that by changing the factor axes, the basic structure of the data, as found by the factor analysis, does not get altered by rotation. The new loadings can often give much better meaning or interpretation to the data. We rotate the factors to simplify the pattern of factor loadings so that they have the properties of 'simple structure'(3). These properties are:

(1) Each row of the factor loading matrix should have at least one zero.

(2) If there are m common factors, each column of the factor loadings matrix should have at least m zeroes.
(3) For every pair of columns of the factor loadings matrix there should be several variables whose entries vanish in one column but not in the other.
(4) For every pair of column of the factor loadings matrix there should be only a number of variables with high loadings in both columns.



line indicates values present.

The two commonly used methods of rotations consistent with simple structure are 'varimax' and 'promax'. The varimax rotation is one of the so-called orthogonal (uncorrelated) rotations, which means that the axes are kept at right angles to each other. On the other hand, promax rotation is an example of oblique (or correlated) rotation where the axes are not only rotated but may be bent towards each other to produce factors which are correlated. With batteries of semantic rating scales, the 'promax' method usually gives a slightly 'tidier', more interpretable solution than the varimax method, but basically the same factor structure(4).

5.4 The applications of factor analysis up to the present time have been mainly in the field of psychology, because the methods were invented by psychologists for dealing with certain of their problems. They are planning experiments employing factor analysis to determine a small number of tests to describe the human mind as completely as possible. The methods of factor analysis have been successfully applied in recent years in varying fields as political science, business and medicine. Mukherjee(5) conducted a factor analysis of fourteen measures of individual coffee preferences (such as, pleasant versus unpleasant flavour, cheap versus expensive taste etc) as a basis for determining what dimension, if any, underlay the original measures, and hence could serve as a better basis for understanding as well as analysising individual preferences. The original fourteen measures were reduced to four underlying factors. Stoetzel(6) obtained three factors based on rankings of various types of liquors (rum, whisky, etc) by a sample of French consumers. After looking at factor loadings, he labelled them as sweetness, price and regional popularity factors, based on his external knowledge about the liquor industry.

By means of factor analysis, the computation of multiple or partial correlations, regression coefficients can be greatly simplified when many variables are involved. This is usually done by identifying likely variables from a much larger set of variables(7). Twedt(8) isolated three variable based on a factor analysis of 19 predictor variables (various aspects of advertisements such as size, colour, layout etc) and the criterion variable of readership. He then used these three variables for predicting readership by doing multiple regression. In a study of household brand proneness(9) for 44 specific grocery products a decision was made to delete the age of husband as an independent variable in the analysis because it had an extremely high correlation with the wife's age. Similarly Massy(10) in studying the variation in the television ownership in 240 urban areas conducted a factor analysis of 27 variables, the percentage distribution of households in 14 income categories, 9 education categories and 4 measures of T.V. coverage. Of these 27 variables, no more than 10 resulting new measures were utilised in subsequent regression analysis.

Factor analysis is also used to obtain factor scores of individuals in the sample for further analysis. This not only reduces the large sets of data to a manageable level but also removes the collinearity in the original variables. Farley(11) used factor scores to explain variability in brand loyalty across products.

REFERENCES

(1) Harman H H 'Modern Factor Analysis' University of Chicago Press Chicago 1967

(2) Lawley D N and Maxwell A E 'Factor Analysis as a Statistical Method' Butterworth London 1971.

(3) Thurstone L L 'Multiple Factor Analysis' The University Chicago Press Chicago 1947

(4) Harris P T 'An Introduction to Multivariate Analysis' Market Research Society London 1973

(5) Mukherjee B N 'A Factor Analysis of some Qualitative Attributes of Coffee' Journal of Advertising Research Vol V No 1 March 1965

(6) Steotzel J 'A Factor Analysis of the Liquor Preferences of French Consumers' Journal of Advertising Research Vol 1 No 2 December 1966 p 7-11

(7) Abraham T P & Hoobakht A 'Application of Factor Analysis for Interpretation of Soil Analysis Data' Indian Journal of Agricultural Statistics Vol XXVI No 1 June 1974

(8) Twedt D W 'A Multiple Factor Analysis of Advertising Readership' Journal of Applied Psychology Vol 36 June 1952

(9) Frank R E & Boyd H W 'Are Private Brand Prone Grocery Customers Really Different?' Journal of Advertising Research Vol V No 4 December 1965

(10) Massy W F 'Television Ownership in 1950, Results of a Factor Analytic Study' In Frank R E Kuehn A A & Massy W F (eds) Quantitative Techniques in Marketing Analysis, Homewood Illinois

(11) Farley J U 'Why does Brand Loyalty vary over products? Journal of Marketing Research Vol 1 November 1964

6. Cluster Analysis

6.1 Cluster analysis is the general procedure by which we objectively group together entities on the basis of their similarities or differences[1]. We can either group together the general properties of the objects called clustering of variables — V analysis[2] or group together the objects into types or classes, called the clustering of objects — O analysis[3].

6.2 The basic criterion for cluster analysis technique is that clusters should be within-group similar and between-group different. There are three main measures of similarity:

i) Distance measures
ii) Correlation measures
iii) Similarity measures designed for attribute data.

6.2.1 Distance Measures: Distance measures tend to be exclusively Euclidean distance D

$$D = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$$

Small D indicates closeness.

Distances among the objects in a given cluster will approach zero but the interspace differences between objects in different clusters will not be zero. Centres of density are located by working with the distance matrix.

Two problems exist with the use of Euclidéan distance measure: (1) Correlated characteristics of the variables and (2) the non-comparability of the original units in which the characteristics are measured. The second is usually 'solved' by standardising all the characteristics to mean zero and unit standard deviation. Thus it is assumed that the mean and variance among characteristics is not important in the grouping process. The first problem can be handled by using principal component analysis (chapter 4) on the characteristics and the factor scores computed for the objects. Each component score may then be weighted by the square root of the eigenvalue associated with that component before computing the distance measure. In practice, distance measure of this kind is usually used when the data are at least intervally scaled.

6.2.2 Correlation measures: The correlation coefficient as a measure of similarity between two objects is widely used. Completely different results are obtained in cluster analysis if one uses the correlation coefficient instead of Euclidean distance measure, and one should attach different interpretation to the results. Distances (D) measure differences between people more powerfully than the correlation measure. Correlations, however, will measure patterns in responses, regardless of the distance between patterns. The correlation coefficient can be considered as a type of distance measure.

If you consider two sectors V and U, the distance between them is given by

$$D = \sqrt{(V-U)' \ (V-U)}$$

$$= \sqrt{(V'V + U'U - 2V'U)}$$

If V and U are scaled to zero mean and unit length, then

$$D = \sqrt{2 - 2r \ .}$$

Three problems are related to this technique. There is loss of information as the correlation removes the elevation and scatter of each object. Some objects may be split among clusters when one uses the factor loadings for grouping objects. Finally, the analyst must usually resort to an variable analysis to interpret the clusters' characteristics according to their correlations with the underlying factors.

6.2.3 Similarity Measures (Attribute data): Similarlity measures are useful when the characteristics of each object are only nominally scaled, which is often the case of attribute data. The usual notion of distance is less applicable here, however, it is still possible using what is known as multi-dimensional scaling (discussed later in chapter 7). The similarity measure in the match coefficient obtained by attribute matching.

If two objects are compared on each of the 10 attributes, with the following results:

	Attribute									
Object	1	2	3	4	5	6	7	8	9	10
1	1	0	1	0	0	1	1	0	1	0
2	0	1	1	0	1	0	1	1	1	1

The fractional match coefficient is given by the number of like entries in the columns, divided by the number of columns, e.g. 4/10. If weak matches (non-possession of attribute) are to be deemphasized, one can use modified match coefficient, using only the possession of attribute as a factor in the ratio, in this example 3/9.

Similarity coefficients have a set of limitations. Firstly, if a group is to be formed on the basis of overall matches, two objects may not be grouped even if they match well on some subset of characteristics. Secondly, if a large number of characteristics are involved, objects which match may do so for accidental reason, reflecting the noise in the data. Thirdly, if some variables are dichotomus and others are multichotomous, the two state attributes will tend to be more heavily weighted in the similarity measures. Finally, if continuous data are discretised in order to use similarity measures, valuable information can be lost. The analyst thus has the problem of deciding the kinds of attributes to include and the number of states to be associated with each.

Scattergram will generally give an idea of grouping, linearity and heteroscedasticity.

6.3 Key-Cluster Factoring: The objective of key-cluster factoring is 1) to select mutually collinear variables defining each of the clusters, 2) On the basis of the proportion of communalities accountable from the scores on the dimensions to select minimal or salient k classes and 3) to provide information when more clusters than minimally sufficient which one to delete.

There are five criteria to keep in mind when assessing the results of the initial factoring procedure:

(1) the degree of collinearity of the definers of each dimension,

(2) the degree of independence of the oblique cluster that defines each dimension

(3) the meaning of each defining oblique cluster as a construcy,

(4) the contribution of the definers to the reliability of a cluster score (which is the sum of Z scores of the variables in the cluster) on the oblique cluster.

and (5) the generality of each oblique cluster and of the variables that define it.

In case of graphical method the nearness of variables is used in selecting the variables. Otherwise one should use the index of collinearity P2 (1, d)

Two variables V_1 and V_2 are said to be perfectly collinear if the correlations variable V_1 has with all other variables (V_3, V_4.....V_n) is a perfect ratio of the correlations V_2 has with all the same variables (V_3, V_4, ...V_n).

If V_1 and V_2 are perfectly collinear

$$\frac{r_{1i}}{r_{2i}} = c \text{ for } i = 3, 4....n$$

$$\text{or } r_{1i} = c\, r_{2i}$$

where r_{1i} and r_{2i} are the correlations variables 1 and 2 have with variables i. Then the Tyron's index of collinearity P2 is

$$P2_{r_1 r_2} = \frac{(\sum_{i=3}^{n} r_{1i} r_{2i})^2}{(\sum_{i=3}^{n} r_{1i}^2)\ (\sum_{i=3}^{n} r_{2i}^2)}$$

$$= I \text{ if perfectly collinear}$$

Thus the closer P2 is to one, the more collinear the variables are.

The cut off rule for finding clusters is, if less than 5% of the estimated overall communalities is accounted by the new cluster, stop finding new clusters. How many clusters to form also depends on the knowledge of the analyst of the particular project and the meaningful interpretation he/she can give to the clusters.

Key clustering being near natural grouping should be used over varimax, principal axes methods, (see Chapters 4 and 5). The graphic solution may indicate the variables of oblique clusters.

To illustrate a feature analysis model of pattern recognition, Lindsay and Norman in their book 'Human Information Processing' analysed the upper-case letters of the alphabet in terms of a likely set of features.

These seven features were:

(1) Vertical lines, (2) Horizontal lines, (3) Oblique lines, (4) Right angles, (5) Oblique angles, (6) Closed curves, (7) Open curves.

The attributes of the letter 'A' were thus one horizontal line, two oblique lines and three oblique angles and the letter 'Q', one oblique line, two oblique angles and one closed curve. Of course, this feature list, used mainly for purposes of illustration, seemed plausible as did many others but usually no evidence was presented to suggest that our visual system did analyse input in the way suggested.

One way of assessing the adequacy of the above feature list would be to carry out a Cluster Analysis to identify similarities and differences amongst letters and to assess the extent to which the resulting clusters made sense. If the clusters seemed reasonable, then this, perhaps, might be regarded as a weak support of the model.

If not, then:

(a) the feature model itself might be inappropriate;
(b) an inappropriate set of features might have been chosen;
(c) the appeal to intuition might not be valid;
(d) clustering might not be a suitable technique for analysing these data.

The Furthest Neighbour Method gives

cluster	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	W	X	Z
V,Y	4	13	5	9	11	9	7	10	10	5	3	6	4	4	5	10	4	9	6	7	6	5	2	5

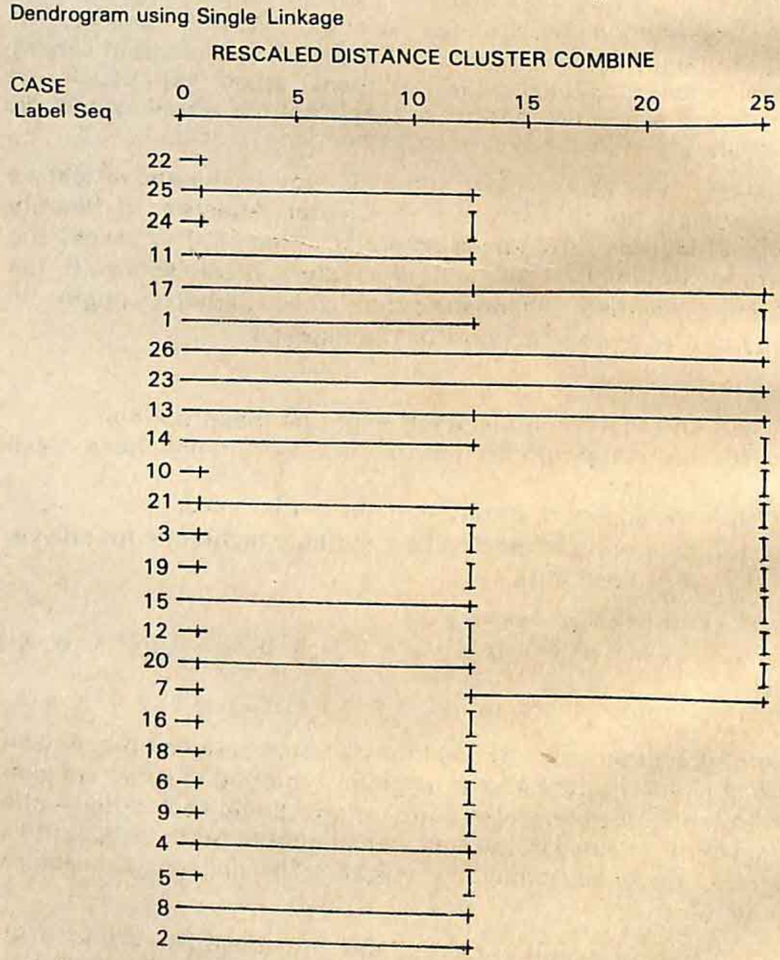
Here, the larger value defines the distance between the remaining clusters. In the nearest neighbour method clusters are joined in terms of the nearest entity in are cluster to another — the single link — while in the furthest neighbour method all entities in one cluster are linked to another. The linkage is therefore complete.

With the seven feature data, the minimum distance or nearest neighbour method runs into problems because the distances at which the larger clusters are combined is the same as the distances at which the smaller one were joined. This method forces the between clusters distances down very rapidly.

The three clusters solution is:

(1) A K Q V W X Y Z
(2) B C D E F G H I J L O P R S T U
(3) M N

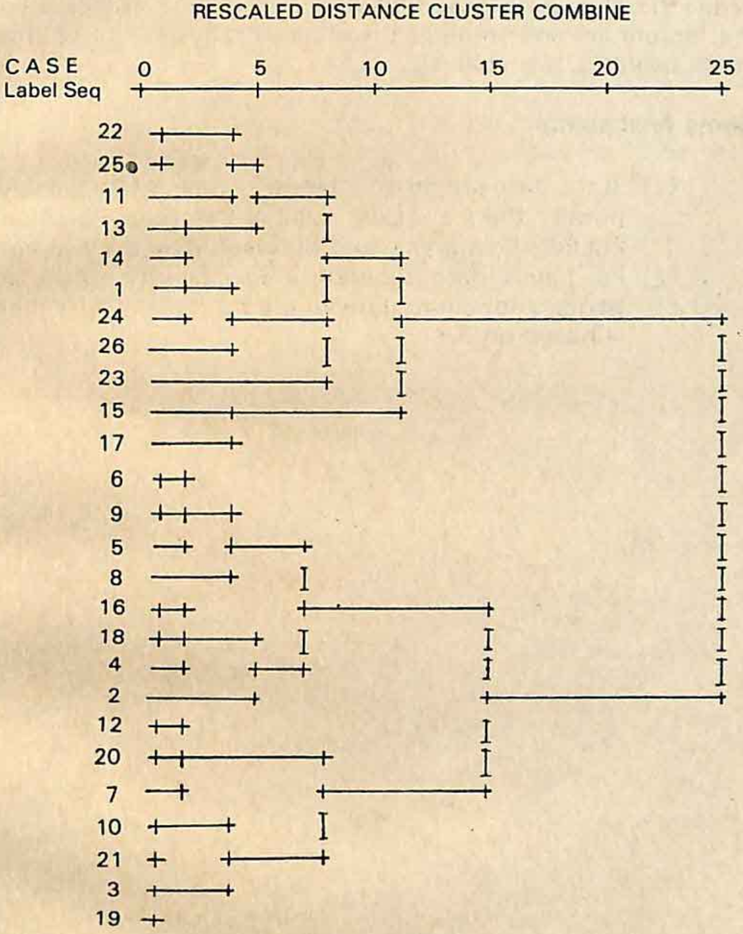
This agglomeration is illustrated in the rather strange looking DENDOGRAM below.



Note that in the dendogram the distance units 1-3 are rescaled to 0-25.

The cluster membership using complete linkage is
(A,K,Q,V,W,X,Y,Z) (B,C,D,E,F,G,H,I,J,L,O,P,R,S,T,U) (M,N)

Dendrogram using Complete Linkage



Some of the furthest neighbour groups seem wrong – a better classification might be

(A,K,M,N,V,W,X,Y,Z) (E,F,H,I,L,T) (B,C,D,G,J,P,R,S,U) (O,Q)

The features – oblique angles and right angles – force a counter intuitive classification. Of course, if it happened that people confused C's and T's more than they confused T's and I's then the furthest neighbour classification might have some merit.

Quite probably,the sum of features not shared by two letters does not provide an adequate distance metric but, on the other hand, the clustering does illustrate both some problems with the feature list and some of the choices that have to be made when using cluster analysis.

Some final points

(1) If the data are on an interval or near interval scale or better, then consider taking the defaults-squared Euclidian distances and between clusters averages.

(2) For binary data generate a dissimilarity matrix and, perhaps for count data where the dissimilarity matrix is based on χ^2.

REFERENCES

(1) For different measures of similarities and classification procedures refer:

- (a) Bignen E J 'Cluster Analysis: Survey & Evaluation of Techniques' Groningen Tilburn University Press 1970
- (b) Sokal P R & Sneath P H A 'Principles of Numerical Taxonomy' W H Freeman & Co 1963
- (c) Joyce T & Channon C 'Classifying Market Survey Respondents' Applied Statistics Royal Statistical Society, London, 1966
- (d) Tyron R C & Bailey D E 'Cluster Analysis' McGraw Hill 1970

(2) Green P E Carmone F J & Fox L B 'Television Programme Similarities: An Application of Subjective Clustering' Journal of the Market Research Society London Vol II No 1 1969

(3) Emmett B P 'The Exploration of Inter-Relationships in Survey Data' Journal of the Market Research Society London Vol 10 No 2 1968
Rothman J & Rauta I 'Towards a typology of the T V audience' Journal of the Market Research Society London Vol 11 No 1 1969

7. Automatic Interaction Detector

Automatic Interaction Detector (AID) is a computer programme which operates under the University of Michigan Executive System and was extensively developed by Morgan and Sonquist (1, 2). It is focussed on a particular kind of data-analysis problem, characteristic of many social science research situations, in which the purpose of the analysis involves more than the reporting of descriptive statistics, but may not necessarily involve the exact testing of specific hypotheses. In this type of situation the problem is often one of determining which of the variables, for which data have been collected, are related to the phenomenon in question, under what conditions, and through what intervening processes, with appropriate controls for spuriousness.

The data-model to which the procedure is applicable may be termed a ''sample survey model'', in which values of a set of predictors X_1, X_2,.....X_n, and a dependent variable Y, have been obtained over a set of observations. In particular this analysis situation is defined to be one in which the X_i are a mixture of nominal and/or ordinal scales and Y is a continuous, or equal interval scale.

7.1 The AID Technique: Regarding one of the variables as a dependent variable, the analysis employs a nonsymmetricl branching process, based on variance analysis techniques, to subdivide the sample into a series of subgroups which maximise one's ability to predict values of the dependent variable. Linearity and additivity assumptions inherent in conventional multiple regression techniques are not required.

In actual operation, the program works as follows:

1. The total input sample is considered the first (and indeed only) group at the start.

2. Select that unsplit sample group, group i, which has the largest total sums of squares.

$$TSS_i = \sum_{a=1}^{N_i} Y^2 - \frac{\left(\sum_{a=1}^{N_i} Y_\alpha \right)^2}{N_i} \qquad (1)$$

such that for the i'th group

$$TSS_i \geq R\,(TSS_T) \text{ and } N_i \geq M \qquad (2)$$

where R is an arbitrary parameter (normally $.01 \leq R \leq .10$) and M is an arbitrary integer (normally $20 \leq S \leq 40$).

The requirement (2) is made to prevent groups with little variation in them, or small number of observations, or both, from being split. That group with the largest total sum of squares (around its own mean) is selected, provided that this quantity is larger than a specified fraction of the original total sum of squares (around the grand mean), and that this group contains more than some minimum number of cases (so that any further splits will be credible and have some sampling stability as well as reducing the error variance in the sample).

3. Find the division of the C_k classes of any single predictor X_k such that combining classes to form the partition p of this group i into two non-overlapping subgroups on this basis provides the largest reduction in the unexplained sum of squares. Thus, choose a partition so as to maximise the expression

$$(n_1 \overline{Y}_1^{\,2} + n_2 Y_2{}^{-2}) - N_i \overline{Y}_i^{\,2} = BSS_{ikp} \qquad (3)$$

where $N_i = n_1 + n_2$

and $\overline{Y}_i = \dfrac{n_1 \overline{Y}_1 + n_2 \overline{Y}_2}{N_i}$

for group i over all possible binary splits on all predictors, with restrictions that (a) the classes of each predictor are ordered into a descending sequence, using their means as a key and (b) observations belonging to classes which are not continuous (after sorting) are not placed together in one of the new groups to be formed. Restriction (a) may be removed, by option, for any predictor X_k.

4. For a partition p on variable k over group i to take place. after the completion of step 3, it is required that

$$BSS_{ikp} \geq Q(TSS_T) \qquad (4)$$

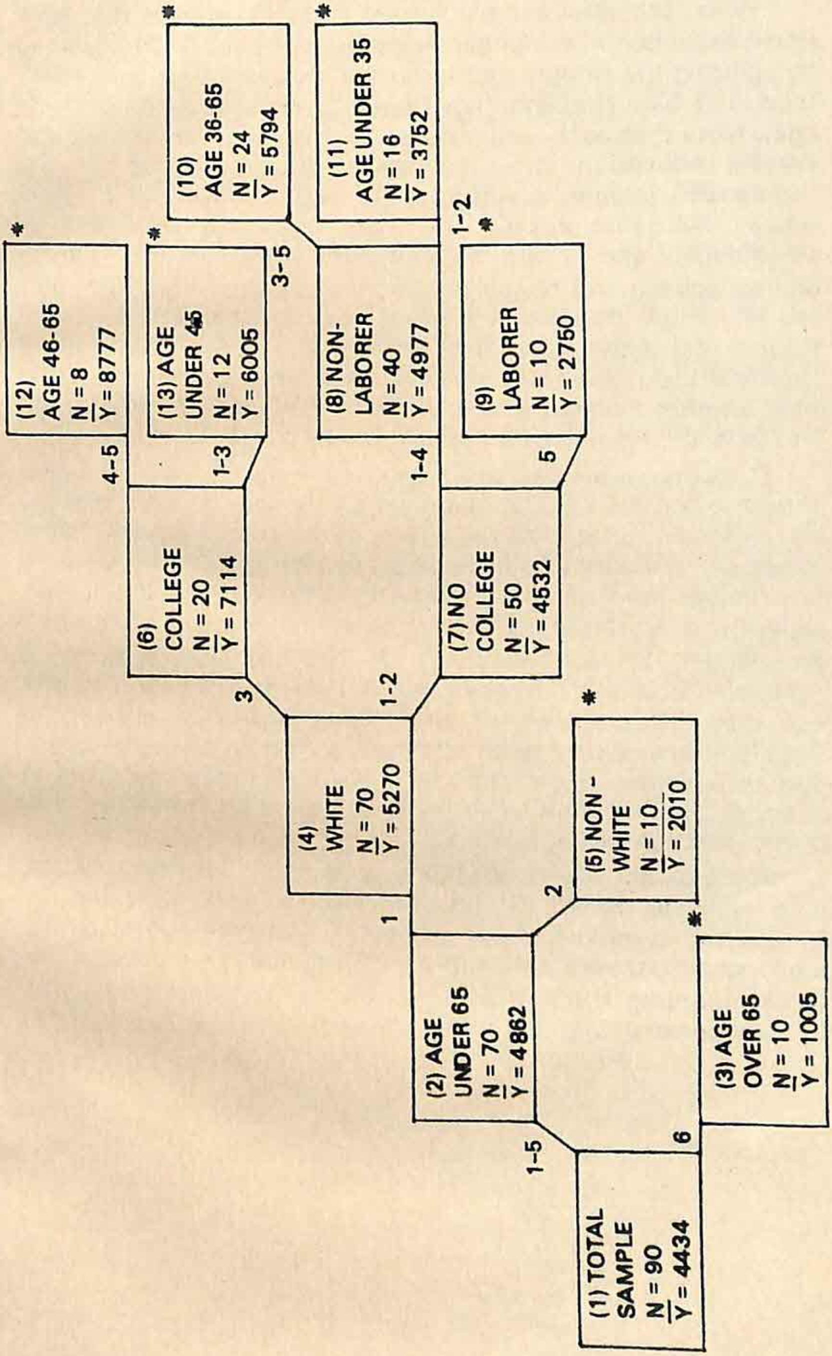
where Q is an arbitrary parameter in the range $.001 \leq Q < R$, and TSS_T is the total sum of squares for the input sample. Otherwise group i is not capable of being split; that is, no variable is "useful" in reducing the predictive error in this group. The next most promising group (TSS_i = maximum) is selected via step 2 and step 3 is then applied to it, etc.

5. If there are no more unsplit groups such that requirement (2) is met, or if, for those groups meeting it, requirement (4) is not met (i.e. there is no "useful" predictor), or if the number of currently unsplit groups exceeds a specified input parameter, the process terminates.

7.2 Examples: In predicting Income (2) Age, Race, Education, Occupation and Length of Time in present jobs are used. Age is an ordered series of categories represented by the numbers (1, 2,...6). Race is coded (1 or 2), Occupation is coded (1, 2,....5), Education is coded (1, 2, 3) and Time on Job is coded (1, 2,...,5). We find the following mutually exclusive groups whose means may be used to predict the income of observations falling into that group:

Group	*Type*	*N*	*Mean Income*	σ
12	Age 46-65, white, college	8	$8777	$773
13	Age under 45, white, college	12	6005	812
10	Age 36-65 white, no college non laborer	24	5794	487
11	Age under 35, white, no college, non-laborer	16	3752	559
9	Age under 65, white, no college, laborer	10	2750	250
5	Age under 65, nonwhite	10	2010	10
3	Age over 65	10	1005	5
Total		90	4434	2263

A one-way analysis of variance over these seven groups would account for 95 per cent of the variation in income. These results are arrived at by the following procedure, as represented by the tree of binary splits:


(1) TOTAL SAMPLE
N = 90
Ȳ = 4434
1–5
6
(2) AGE UNDER 65
N = 70
Ȳ = 4862
(3) AGE OVER 65
N = 10
Ȳ = 1005
*
1
2
(4) WHITE
N = 70
Ȳ = 5270
(5) NON - WHITE
N = 10
Ȳ = 2010
*
3
1-2
(6) COLLEGE
N = 20
Ȳ = 7114
(7) NO COLLEGE
N = 50
Ȳ = 4532
4–5
1-3
(12) AGE 46-65
N = 8
Ȳ = 8777
*
(13) AGE UNDER 45
N = 12
Ȳ = 6005
*
1-4
5
(8) NON-LABORER
N = 40
Ȳ = 4977
(9) LABORER
N = 10
Ȳ = 2750
*
3–5
1–2
(10) AGE 36-65
N = 24
Ȳ = 5794
*
(11) AGE UNDER 35
N = 16
Ȳ = 3752
*

When the total sample (group 1) is examined, the maximum reduction in the unexplained sum of squares is obtained by splitting the sample into two new groups, ''age under 65'' (classes 1-5 on age) and ''age 65 and over'') (those coded 6 on age). Note that each group may contain some nonwhites and varying education and occupation. groups. Group 2, the ''under-65' people, are then split into ''white'' and ''nonwhite''. Note that group 5, the ''nonwhites'', are all under age 65. Similarly, the ''white, under age 65'' group is further divided into college and noncollege individuals etc. A group which can no longer be split is marked with an asterisk and constitutes one of the above final groups. The variable ''Length of Time in Present Job'' has not been used. At each step there existed another variable which proved more useful in explaining the variance remaining in that particular group.

7.3 Limitations and Applications: There are several limitations in using AID (3) Data sets with a thousand cases or more are necessary, otherwise the power of the search process must be restricted drastically or those processes will carry one into a never-never land of idiosyncratic results (4) A well-behaved dependent variable without extreme cases of severe bimodalities is also assumed. A dichotomous dependent ''variable'' is usable if it takes on of its values more than 20 and less than 80 per cent of time. The predictors should be classifications, where each of the classes is in a single dimension; otherwise one really should make dichotomies out of each of the categories. Finally, some theory must be applied, if only in the selection of the predictors.

In the recent years, AID has been used in marketing. In one such analysis, Heald (5) has studied the factors which influence the turnover of the outlets. The technique also has been used in assessing the store performance and site selection (6), segmenting the markets by Group Purchasing behaviour (7), and constructing marketing indicators (8).

REFERENCES

(1) Morgan J N & Sonquist J A "Problems in the Analysis of Survey Data and a Proposal" Journal of the American Statistical Association 58 June 1963 415-435

(2) Morgan J N & Sonquist J A 'The Determination of Interaction Effects' Monograph No. 35 Institute of Social Research University of Michigan 1964

(3) Doyle P 'The use of AID and Similar Research Procedures' O R Quarterly September 1973

(4) Doyel P & Fenwick I 'The Pitfalls of AID Analysis' Journal of Marketing Research Vol XII November 1975

(5) Heald G I 'New Approach to Retail Research' Seventh Annual Conference The Market Research Society March 1974

(6) Chib R & Kidgell J 'A New Approach to Assessment of Store Performance & Site Selection' (Memeio-graph) Gallup Poll London

(7) Assael H 'Segmenting Markets by Group Purchasing Behaviour: An application of the AID Technique' Journal of Marketing Research Vol VII May 1970 153-8

(8) Barnes W N 'International Marketing Indicators' Monograph European Journal of Marketing July 1979

8. Non-Metric Multi-Dimensional Scaling (NMS)

Numerous qualitative approaches (1) are used to study the attitude of potential consumers to brands and advertising campaigns. The technique of NMS is simple.

The basic idea of the method (2) is that one can determine how consumers view competing products or brands without asking them complicated questions requiring numerical ratings along various scales deemed of importance. Instead a virtually complete picture can be built up by asking respondents simply to rank pairs of products in the order of overall similarity.

For example, a respondent would be given a pile of cards. On each card would be the names (or descriptions) of two products from the assortment under consideration. She is then shown how to arrange the cards in her order of decreasing similarity, so that the pair at the bottom will be the least similar and the pair at the top the most alike. These ordered cards are then fed into the programme and the resulting output is a configuration of points (products) in multidimensional (usually two or three) space. The distances betwen the products should

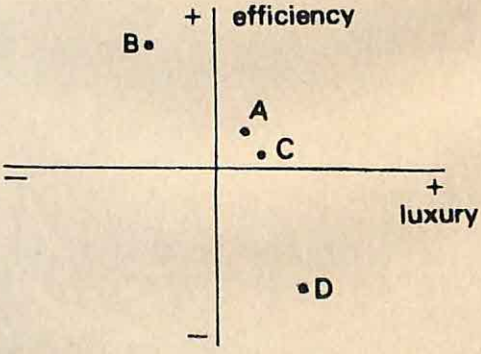


Figure 1

then be consistent with the rankings given by the respondent. Thus if four brands A, B, C, D are ranked so that A and C are the most similar and B and D the least, the configuration might look like figure 1. Furthermore, the dimensions will usually be interpretable. Here B is shown to be seen as an 'efficient' product whereas D is is low on 'efficiency' but seen as high on the 'luxury' dimension.

Spatial configurations are reproduced from only ranked data and the ranks act as constraints which greatly limit where the alogrithm can place the points.

8.1 Advantages of NMS: Compared to the traditional methods, NMS has some appealing features (3)

(1) Only ranked data are required.
— Give more precise information without sacrificing usability

(2) Factors are not pre-specified
— No pre-structuring of questions

(3) Considers all relevant dimensions
— Unlike most attitude scaling treats perceptions of consumers as multi-dimensional

(4) Differences in views:
— Allows segmenting the respondents

(5) Incorporating Preferences:
— Incorporates perception with preference

Finally,

(6) Handling missing data:
— Allows sighting missing data without biasing other relationships
— Remaining data sufficient to constrain the configuration.

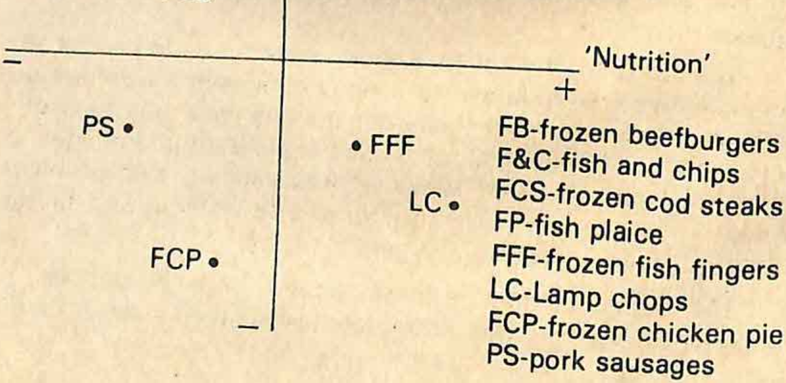
8.2 Application of the Technique: NMS is successfully used in market segmentation analysis (2,3) and new product studies and also for test marketing, developing perceptual spaces (4).

In a study for market for convenience foods in the UK (5) the objective was to know how many dimensions were important to consumers, what these dimensions were and how different products were evaluated against each other and against their preferences. Other ways of approaching this problem would have been through structured interviewing and factor analysing the results and also gap analysis.

The essential data were collected from a sample of housewives in the target group. Each was given a set of cards containing the names of all pairs of eight products under study and shown how to sort them in the order of decreasing overall similarity. How similarity was defined was left open to the respondent. Next, the housewives ordered the brands in terms of preference. Finally, information was collected about usage and attitudes from a conventional questionaire.

First, two points of view were distinguished — one group being heavy users and the other light users. Both groups were then analysed using the NMS programme. Three dimensions appeared to satisfactorily describe the perceptions of consumers; these dimensions appeared to be 'nutrition' 'value for money' and 'substantiality'. The relative position of the products for the heavy user group using two of the dimensions is shown in the diagram. Finally, the preference data to calculate an average 'ideal point' was used.

However like all other techniques there are limitations to use of NMS. First stems from the computational side. How unique are the attribute spaces given noisy and/or missing data? How reliable, statistically, are the solutions? Green (5) and Percy (6) have studied this problem and have suggested certain improvements to tackle it. Secondly, there is the problem concerning distance measurement. Secondly, there is the problem concerning distance measurement. There are other distance measures, apart from the Euclidean distance. These measures may give drastically different interpretations. However, NMS used with proper discrimination offers a possibility of new insights into analysis of market behaviour.



REFERENCES

(1) Shepard R N 'The Analysis of Proximities: Multidimensional Scaling With an Unknown Distance Function' Part One Psychometrika Vol 27 1962 125-139

(2) Doyle P & Hutchinson P 'Measuring Consumer Needs and Beliefs' ADMAP October 1972

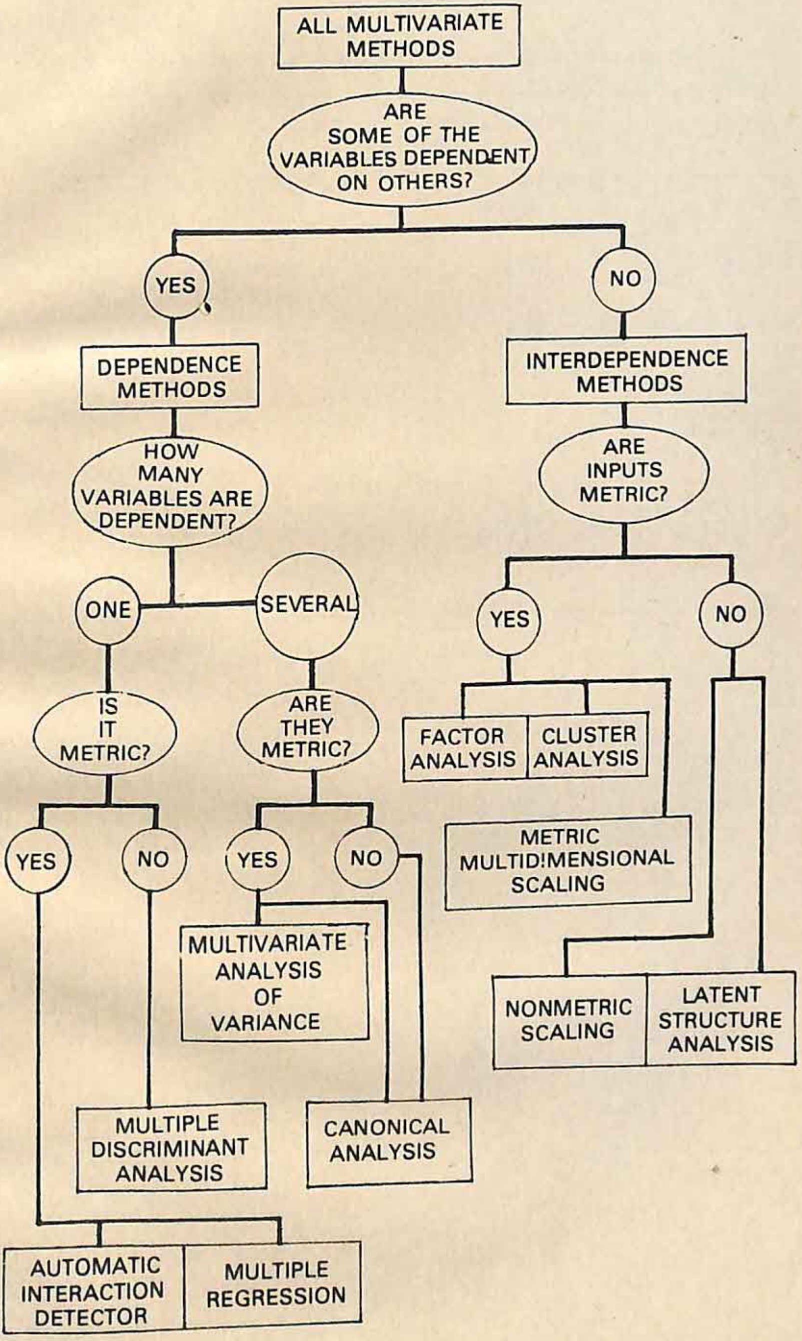
(3) Doyle P & McGee J 'Alternative Convenience Foods' Journal of the Market Research Society Vol 15 No 1 January 1973

(4) Neidell L A 'The Use of Nonmetric Multidimensional Scaling in Marketing Analysis' Journal of Marketing Vol 38 October 1969 37-43

(5) Green P E 'On the Robustness of Multidimensional Scaling Techniques' Journal of Marketing Research Vol XII February 1975 73-81

(6) Percy L H 'Multidimensional Unfolding of Profile Data: A Discussion & Illustration with Attention to Badness-of-fit' Journal of Marketing Research Vol XII February 1975 93-99

APPENDIX A


ALL MULTIVARIATE METHODS
ARE SOME OF THE VARIABLES DEPENDENT ON OTHERS?
YES
NO
DEPENDENCE METHODS
INTERDEPENDENCE METHODS
HOW MANY VARIABLES ARE DEPENDENT?
ARE INPUTS METRIC?
ONE
SEVERAL
YES
NO
IS IT METRIC?
ARE THEY METRIC?
FACTOR ANALYSIS
CLUSTER ANALYSIS
YES
NO
YES
NO
METRIC MULTID!MENSIONAL SCALING
MULTIVARIATE ANALYSIS OF VARIANCE
NONMETRIC SCALING
LATENT STRUCTURE ANALYSIS
MULTIPLE DISCRIMINANT ANALYSIS
CANONICAL ANALYSIS
AUTOMATIC INTERACTION DETECTOR
MULTIPLE REGRESSION

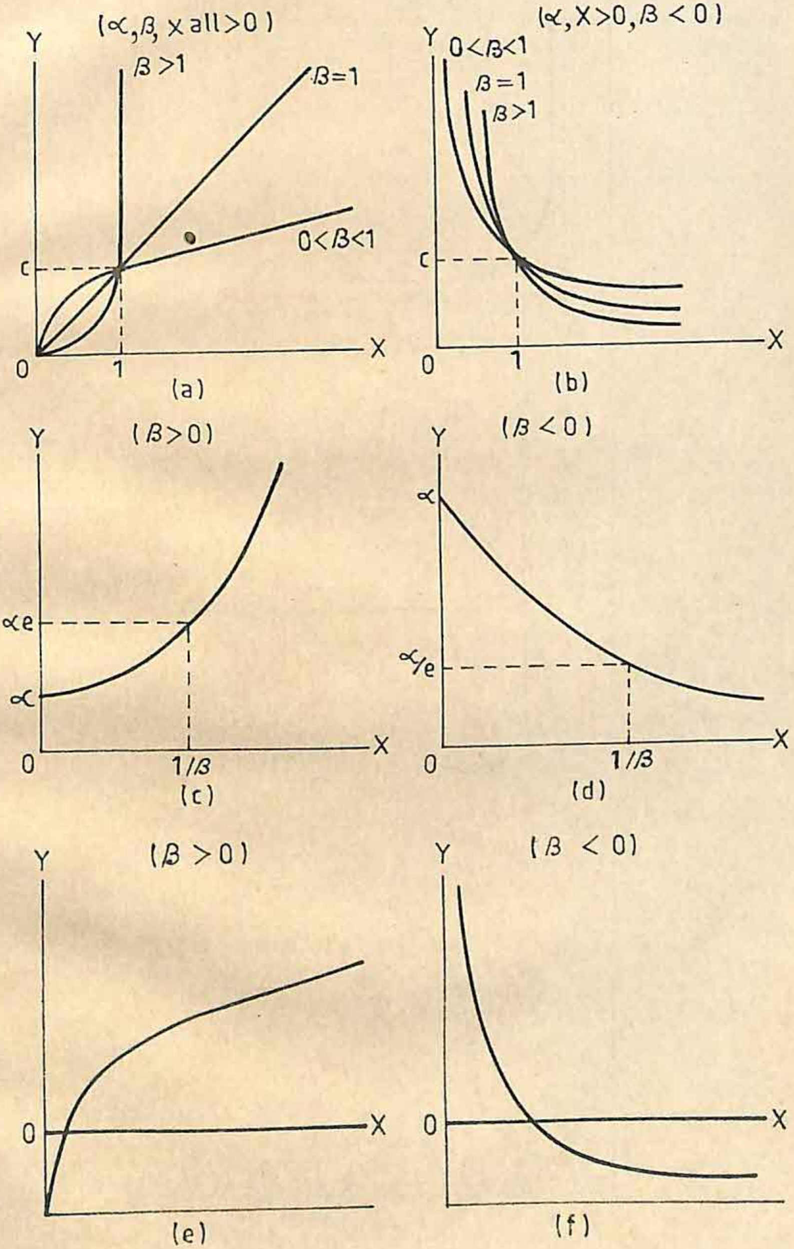
APPENDIX B

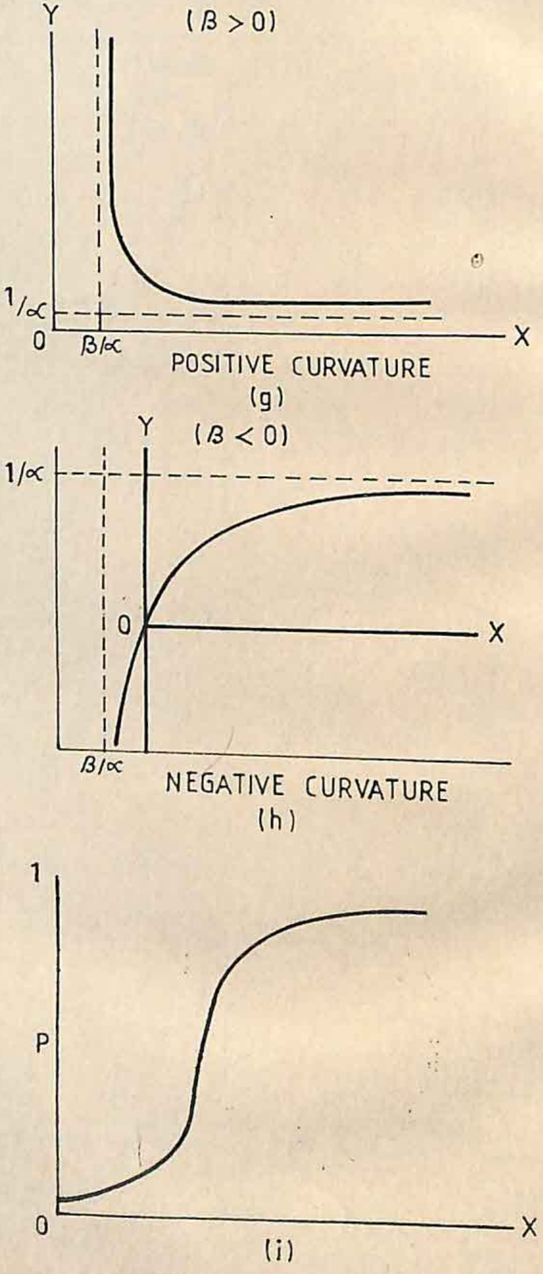
Linearizable functions with corresponding transformations

Function	*Transformation*	*Linear form*	*Graph shown in figure*
$y = \alpha x^{\beta}$	$y' = \log y,\ x' = \log x$	$y' = \log\alpha + \beta x'$	a,b
$y = \alpha e^{\beta x}$	$y' = 1ny$	$y' = 1n\alpha + \beta x$	c,d
$y = \alpha + \beta \log x$	$x' = \log x$	$y = \alpha + \beta x'$	e,f
$y = \dfrac{x}{\alpha x - \beta}$	$y' = \dfrac{1}{y},\ x' = \dfrac{1}{x}$	$y' = \alpha - \beta x'$	g,h
$y = \dfrac{e^{\alpha+\beta x}}{\alpha+\beta x}$ $1+e$	$y' = 1n\left(\dfrac{y}{1-y}\right)$	$y' = \alpha + \beta x$	i

APPENDIX B

Graphs of Linearizeable Functions





POSITIVE CURVATURE
(g)

NEGATIVE CURVATURE
(h)

(i)

APPENDIX C

Transformations to Stabilise Variance

Probability distribution of variable y	*Variance* y *in terms of its mean* μ	*Transformation*	*Resulting variance*
Poisson	μ	$\sqrt{y}$ or $(\sqrt{y}+\sqrt{y}+1)$	0.25
Binomial	$\frac{\mu(1-\mu)}{n}$	$\text{Sin}^{-1}\sqrt{y}$ (degrees)	$\frac{821}{n}$
		$\text{Sin}^{-1}\sqrt{y}$ (radians)	$\frac{0.25}{n}$
Negative Binomial	$\mu+\lambda^2 m^2$	$\lambda^{-1} s\text{Sinh}^{-1}(\lambda\sqrt{y})$ or $\lambda^{-1}\text{Sinh}^{-1}(\lambda\sqrt{y}+0.5)$	0.25

ABOUT THE AUTHOR

Dr. B.A. Chansarkar completed M.Sc. Statistics in 1962 and M.A. Economics in 1964 from Nagpur University. He was awarded the degree of Doctor of Philosophy from Nagpur University in 1976. He has a wide professional experience. He has worked for six years as Senior Market Research Executive, Watney Mann Ltd., London before joining the Middlesex Polytechnic, Enfield. He has been a resource expert to the Government of Maharashtra and a Visiting Professor in the Nagpur University during 1980-81.

Dr. Chansarkar is a Fellow of The Royal Economics Society, Royal Statistical Society, London, and Econometrics Society, USA. He has published several papers on economics of agriculture, immigration in the UK, education and statistics.

45.